SCALABLE COMPUTING
SOFTWARE LABORATORY

ILLINOIS INSTITUTE
OF TECHNOLOGY

Kougkas A, Devarajan H, Sun XH. I/O acceleration via multi-tiered data buffering and prefetching. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(1): 92–120 Jan. 2020. DOI 10.1007/s11390-020-9781-1

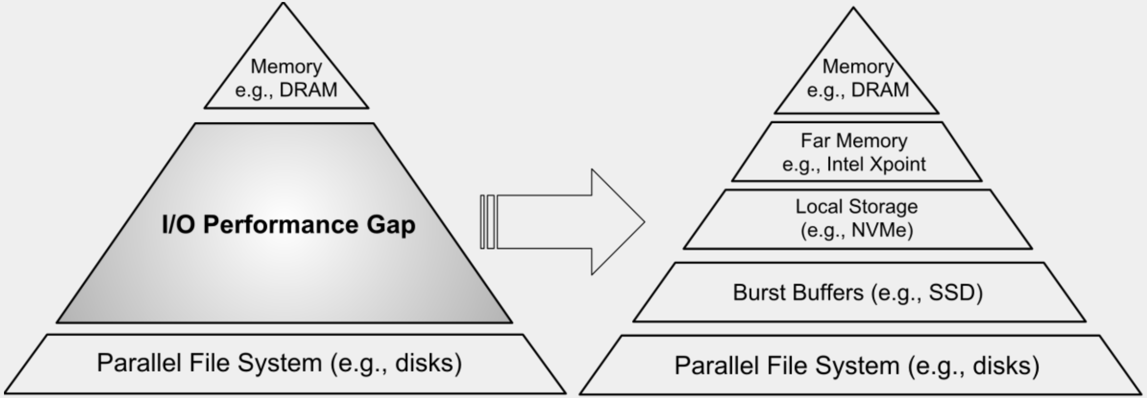
Hermes

A Multi-Tiered Distributed I/O Buffering System

Anthony Kougkas, Hariharan Devarajan , and Xian-He Sun
akougkas@iit.edu, hdevarajan@hawk.iit.edu, sun@iit.edu

Deep Memory and Storage Hierarchy (DMSH)

- New storage system designs incorporate non-volatile burst buffers between the main memory and the disks.
- HPC hierarchical storage systems with burst buffers (BB) have been installed at several HPC sites.
- Multiple levels of memory and storage in a hierarchy, called **DMSH**.



Memory e.g., DRAM
I/O Performance Gap
Parallel File System (e.g., disks)
Memory e.g., DRAM
Far Memory e.g., Intel Xpoint
Local Storage (e.g., NVMe)
Burst Buffers (e.g., SSD)
Parallel File System (e.g., disks)

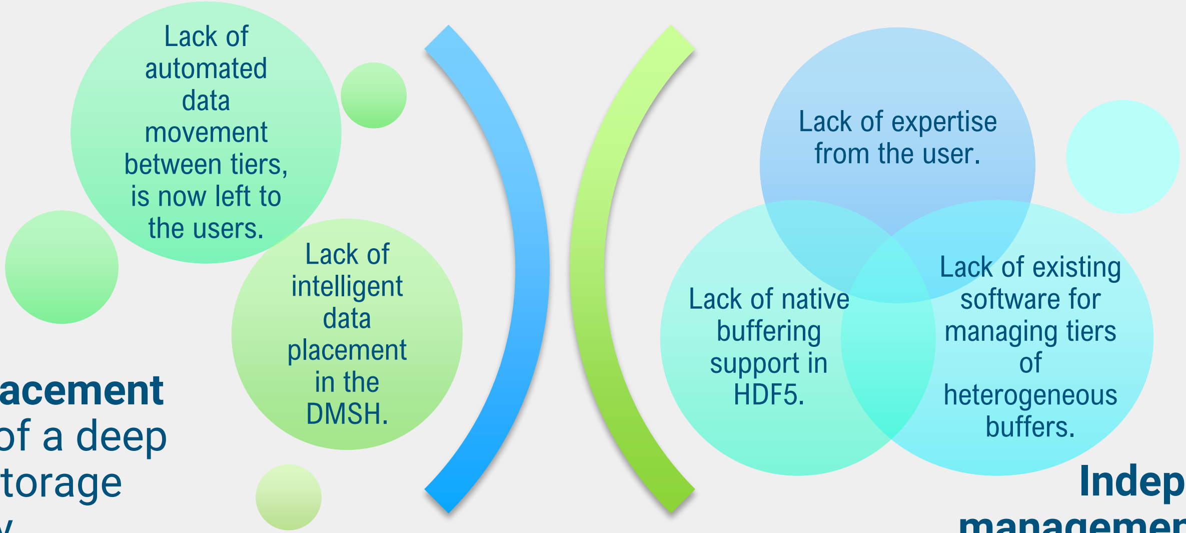
Ideally, the presence of multiple tiers of storage should be **transparent** to applications without having to sacrifice **I/O performance**.

DMSH systems require:

- efficient and transparent **data movement** through the hierarchy
- new data placement algorithms,
- effective memory and metadata management,
- an efficient communication fabric.

Anthony Kougkas, PhD
akougkas@iit.edu

Multi-tiered systems today



Lack of automated data movement between tiers, is now left to the users.

Lack of intelligent data placement in the DMSH.

Complex data placement among the tiers of a deep memory and storage hierarchy

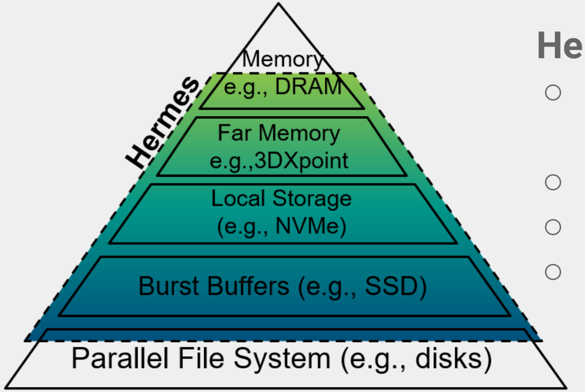
Lack of expertise from the user.

Lack of native buffering support in HDF5.

Lack of existing software for managing tiers of heterogeneous buffers.

Independent management of each tier of the DMSH

Hermes in a snapshot



Hermes is a new, multi-tiered, distributed buffering platform that:

- Enables, manages, and supervises I/O operations in the Deep Memory and Storage Hierarchy (DMSH).
- Offers selective and dynamic layered data placement.
- Is modular, extensible, and performance-oriented.
- Supports a wide variety of applications (scientific, BigData, etc.,).

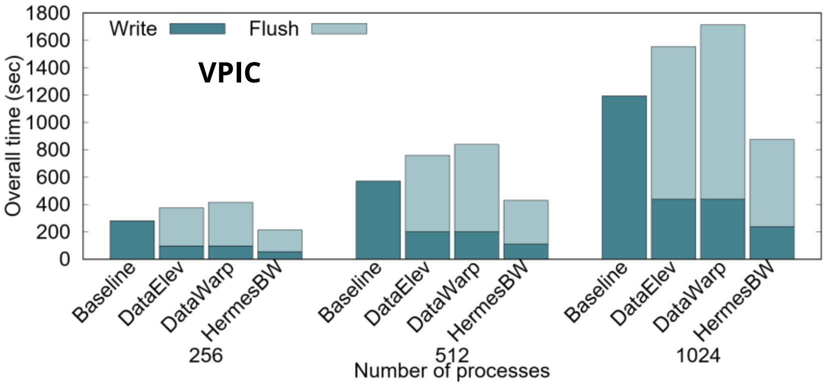
Hermes goals

- being application- and system-aware
- maximizing productivity
- increasing resource utilization
- abstracting data movement
- maximizing performance
- supporting a wide range of scientific applications and domains

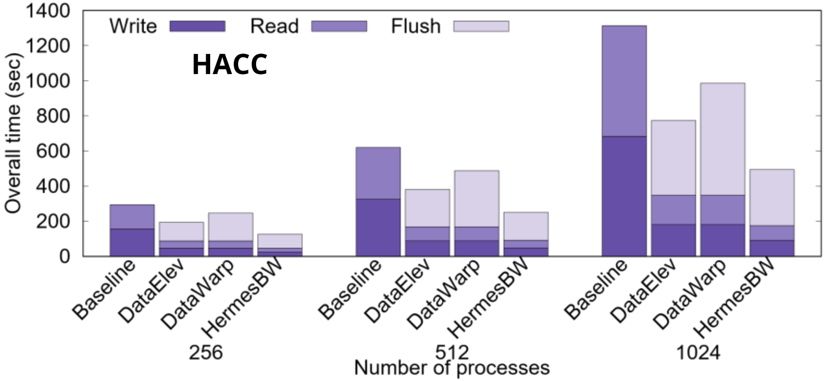
Anthony Kougkas, PhD
akougkas@iit.edu

Evaluation Results

- Vector Particle-In-Cell (VPIC):
  - Uses HDF5 files
- Hardware Accelerated Cosmology Code (HACC):
  - MPI - I/O Independent
- Strong scaled up to 1024 ranks
- 16-time steps
- Metric:
  - Total I/O time (write + read + flush)



Hermes offers **5x and 2x** higher write performance on average when compared to No Buffering and state-of-the-art buffering platforms



Hermes offers **7.5x and 2x** higher read performance for repetitive patterns when compared to No Buffering and state-of-the-art buffering platforms

- Hermes hides data movement between tiers behind compute
- Hermes leverages the extra layers of the DMSH to offer higher BW
- Hermes utilizes a concurrent flushing overlapped with compute

Hermes Contributions

Advance

the state-of-the-art high level I/O libraries with new buffering algorithms and mechanisms that address the challenges of a DMSH system.

Facilitate

an agile architecture that will allow the evolution of next generation I/O and will address the increasingly challenging scale and complexity of future systems.

Enhance

the HDF5 core library with intelligent buffering, caching, and prefetching techniques based on machine learning algorithms.

Support

new scientific and engineering methodologies and computational requirements allowing applications to immediately take advantage of DMSH.